Final Project

Christophe Hunt

May 13, 2017

Contents

Pick one of the quantitative independent variables from the training data set (train.csv), and define that variable as X.

Pick SalePrice as the dependent variable, and define it as Y for the next analysis.

```
library(tidyverse)
train.df <- as_tibble(read.csv("https://raw.githubusercontent.com/ChristopheHunt/MSDA---Coursework/maste
```

```
sub.train.df <- train.df[, c("SalePrice", "LotArea")]
```

> The variable we will set to X is LotArea, which is defined as the Lot size in square feet. I chose this because an anecdotal assumption is that the larger the lot size is the higher the sale price. However, living in NYC there are tiny lots in very desirable places that have a high price so I believe there may be some interesting patterns here.

1 Probability

Calculate as a minimum the below probabilities a through c.

Assume the small letter "x" is estimated as the 4th quartile of the X variable, and the small letter "y" is estimated as the 2nd quartile of the Y variable. Interpret the meaning of all probabilities.

a. $P(X > x|Y > y)$

```
prob.x <- list(qrtx = as.numeric(quantile(sub.train.df$LotArea)[4]),
               mean = mean(sub.train.df$LotArea),
               std = sd(sub.train.df$LotArea))

prob.y <- list(qrtx = as.numeric(quantile(sub.train.df$SalePrice)[2]),
               mean = mean(sub.train.df$SalePrice),
               std = sd(sub.train.df$SalePrice))
```

b. $P(X > x, Y > y)$

c. $P(X < x|Y > y)$

Does splitting the training data in this fashion make them independent?

In other words, does $P(X|Y) = P(X)P(Y))$?

Check mathematically, and then evaluate by running a Chi Square test for association.

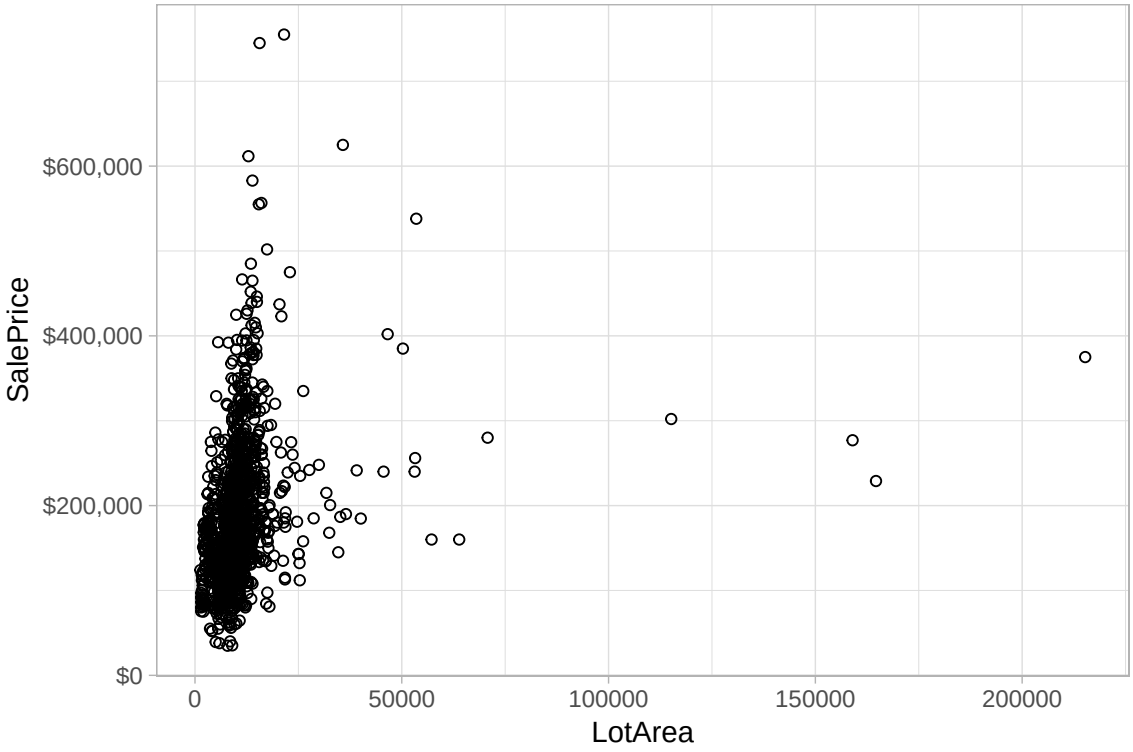
You might have to research this.

2 Descriptive and Inferential Statistics.

Provide univariate descriptive statistics and appropriate plots for both variables.

Provide a scatterplot of X and Y.

```
ggplot(sub.train.df, aes(x = LotArea, y = SalePrice)) + geom_point(shape = 1) +
    theme_light() + scale_y_continuous(labels = dollar)
```



Transform both variables simultaneously using Box-Cox transformations.

> I am using the `BoxCox.lambda` function from the `forecast` package to determine the necessary transformations for the two variables.

λ	Variables
-0.3308	SalePrice
-0.1268	LotArea

Common Box-Cox Transformations[1] [2]

λ	Y'
-0.5	$Y^{-0.5} = \frac{1}{\sqrt{(Y)}}$
0	$\log(Y)$
.25	$\sqrt[4]{Y}$

Lambda values were truncated to the nearest tenth that match a common transformation as per the below table.

variable	variable transformation
SalePrice	$SalePrice^{-0.5}$
LotArea	$log(LotArea)$

Using the transformed variables, run a correlation analysis and interpret.

```
sub.train.df.trans <- sub.train.df %>%
                      mutate(SalePrice = SalePrice^(-.5),
                             LotArea = log(LotArea))

sub.train.cor <- cor.test(sub.train.df$SalePrice,
                          sub.train.df$LotArea,
                          method = "pearson", conf.level = .99)
sub.train.cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  sub.train.df$SalePrice and sub.train.df$LotArea
## t = 10.445, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  0.2000196 0.3254375
## sample estimates:
##       cor
## 0.2638434
```

> The p-value of the correlation test is 2.2e-16 which is less than the significance level of alpha at .05. We are using the standard alpha as there is no indication another any other value for alpha should be used. We can therefore say that lot size and sale price are significantly correlated with a correlation coefficient of 0.264.

Test the hypothesis that the correlation between these variables is 0 and provide a 99% confidence interval.

> The correlation test has specifically done that for us and we can safely reject the null hypothesis as we see that our 99% confidence interval exists at the values (0.200, 0.325) with a p-value < 2.2e-16.

Discuss the meaning of your analysis.

> This means two possible things could have occured, there is no correlation and this data set is pulled from an unusual population of house sales. Or, more likely with the values obtained, our assumption of 0 correlation is incorect and we have obtained a very typical data set and must reject the null hypothesis.

[1] Osborne, Jason W. "Improving your data transformations: Applying the Box-Cox transformation." Practical Assessment, Research & Evaluation 15.12 (2010): 1-9.

[2] By Understanding Both the Concept of Transformation and the Box-Cox Method, Practitioners Will Be Better Prepared to Work with Non-normal Data. . "Making Data Normal Using Box-Cox Power Transformation." ISixSigma. N.p., n.d. Web. 29 Oct. 2016.

3 Linear Algebra and Correlation.

Invert your correlation matrix.(This is known as the precision matrix and contains variance inflation factors on the diagonal.) Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix.

4 Calculus-Based Probability & Statistics

Many times, it makes sense to fit a closed form distribution to data. For your non-transformed independent variable, location shift it so that the minimum value is above zero.

Then load the MASS package and run fitdistr to fit a density function of your choice. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html).

Find the optimal value of the parameters for this distribution, and then take 1000 samples from this distribution (e.g., rexp(1000) for an exponential).

Plot a histogram and compare it with a histogram of your non-transformed original variable.

5 Modeling

Build some type of regression model and submit your model to the competition board.

Provide your complete model summary and results with analysis.

Report your Kaggle.com user name and score.

Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix.